J. INDIZACIÓN Y RECUPERACIÓN DE INFORMACIÓN

Informe de situación

Tendencias en recuperación de información: principios y retos para una nueva década de datos enlazados

Por **Eva Méndez**

Méndez, Eva. "Tendencias en recuperación de información: principios y retos para una nueva década de datos enlazados". *Anuario ThinkEPI*, 2010, v. 4, pp. 231-239

Resumen: *Se presenta un estado de la cuestión en el ámbito de la recuperación de información, haciendo especial hincapié en las últimas tendencias, estrategias y experiencias que han tenido lugar en 2009, así como en las expectativas de investigación y desarrollo en este campo en un futuro próximo (2010). Se recoge en primer lugar una revisión sobre la importancia y complejidad creciente de la recuperación de información en nuestra disciplina, profundizando posteriormente en el reto que tiene la biblioteconomía-documentación de la segunda década del siglo XXI en este sentido: no sólo recuperar información, sino también "enlazar datos".*

Palabras clave: *Recuperación de información, Metadatos, Semántica, Web semántica, Web 2.0, Etiquetado social, Web 3.0, Estándares, Linked data, Portabilidad de datos, Nuevas estrategias de búsqueda.*

Title: ***Trends in information retrieval: principles and challenges for a new decade of linked data***

Abstract: *This document shows a state-of-the-art in the information retrieval landscape, with special emphasis on recent trends, strategies and experiences during 2009, as well as research and development expectations in this field for the near future (2010). The paper first gathers a simple review about the growing importance and complexity of Information Retrieval in our discipline, deepening then about the challenge for Librarianship and Information Science for the second decade of 21st century: not only to retrieve information but also to "link data".*

Keywords: *Information retrieval, Metadata, Semantics, Semantic Web, Web 2.0, Social tagging, Web 3.0, Standards, Linked data, Data portability, New searching strategies*

1. Introducción

LA RECUPERACIÓN DE INFORMACIÓN (*information retrieval*, o simplemente *IR*), es uno de los campos que más interés, investigación y desarrollo experimentó en el siglo XX en el marco de nuestra disciplina.

Además de los estudios tradicionales sobre recuperación de información (**Baeza** et al., **Chowdhury**, entre otros muchos), bases de datos, indización, etc., el siglo XX finalizó con una extensa colección de análisis sobre recuperación de información en internet (RII), sobre el funcionamiento de motores de búsqueda, directorios, portales y bibliotecas digitales recién llegadas al panorama documental; y con ellos, también, nuevas tecnologías para la descripción de la información digital y su recuperación cualificada, como los metadatos, así como nuevas estructuras, lenguajes y vocabularios para organizar la información, combinando tendencias algorítmicas y semánticas para enfrentar, como diría **Weinberger** (2007), el nuevo desorden digital.

Han pasado más de diez años desde que *BackRub*, el proyecto de **Larry Page** y **Sergey Brin**, se convirtiera en *Google*, y un poco menos, desde que éste se convirtiera en "el buscador", eliminando poco a poco de nuestras pantallas a *Altavista*, *Lycos* o *Excite*, que habían hecho las delicias de la búsqueda en internet hasta entonces.

En esta década de tendencia *Google* –que algunos autores como **Dempsey** o **Bawden** denominaron *Amazoogle* o *Googlezon* para reflejar el influjo de los servicios *Amazon* y *Google*– la RII

de carácter global ha pasado a ser una caja de búsqueda y una interfaz sencilla para mostrar los resultados. De esta manera queda transparente para el usuario la evolución de los más sofisticados algoritmos y experimentos que, bien basados en modelos tradicionales de recuperación de información (booleano, espacio-vectorial, etc.) o bien utilizando procesamiento del lenguaje natural (*NLP*) o tecnologías de la web semántica, han tratado de superar a *Google* y también de visibilizar, cada vez más, la web profunda o invisible.

"Los motores de búsqueda se han ganado la confianza de todos, haciéndose indispensables para encontrar el conocimiento en la Web"

Mientras tanto, el desarrollo de portales bibliográficos, bibliotecas digitales y repositorios de toda índole ha ido incorporando servicios de búsqueda dirigida, basados en nuevas ontologías o renovados tesauros, así como en modelos de codificación de metainformación.

Además, en los últimos diez años no sólo ha cambiado la naturaleza de la información que se busca (y se encuentra, en el mejor de los casos) utilizando como interlocutor un navegador web o agente de usuario: ha cambiado también la forma en que se estructura esa información y, por supuesto, ha cambiado la cantidad de información disponible en la Red, así como las herramientas y estándares para generarla y visualizarla. Los urls no sólo sirven para identificar lo que genéricamente llamamos recursos de información, sirven para identificar "cosas conceptuales" (**Berners-Lee**, 2009): personas, eventos, productos, en definitiva, datos. Estamos ahora en un proceso de transición (ampliamente anunciado en esta última década) de una "web de documentos", estructurados en html tradicional o xml, a una "web de datos", codificados fundamentalmente en rdf, pero también, en microformatos xhtml, rdfa y en un futuro aún en construcción, en html 5[5].

Estamos ante un nuevo ecosistema informativo. Lo que **Gruber** llamó hace un par de años sistema de conocimiento colectivo, o simplemente web 3.0, donde la web social y las tecnologías de la web semántica se abrazan para reflejar un nuevo paradigma tecnológico para el procesamiento y gestión de la información. Todo esto afecta a qué y cómo buscamos y plantea un nuevo panorama profesional en la organización y recuperación de información, que tratamos en este pequeño artículo.

2. Buscadores 2009: diez años después de *Google*

Primero se habló de orugas web *(webcrawlers)*, luego, en consonancia con la tela de araña mundial (www), se habló de arañas *(webspiders)*. A estas denominaciones, metafóricas y sin duda curiosas para un mundo tradicional como el de la documentación, se fueron añadiendo otras que pronto adoptarían legos y profanos en la materia, para designar a aquellas herramientas que de forma automática "recorren" la Web, indizándola para su posterior recuperación.

La metáfora continuó y **Witten** (2007) llamó a los motores de búsqueda "dragones web" porque son los guardianes del tesoro de la información de nuestra sociedad. Misteriosos, míticos, mágicos, poderosos, independientes e impredecibles como los dragones, los motores de búsqueda se han ganado la confianza de todos, haciéndose indispensables para encontrar el conocimiento registrado en la Web como medio de información, pero también como medio de comunicación de conocimiento colectivo.

Aunque 2009 fue el año del búfalo en la tradición china, en el entorno de la recuperación de información podríamos considerarlo el año de los dragones de búsqueda. Pero, sobre todo, 2009 ha sido el año de *linked open data* (*LOD*, o simplemente *linked data*). Una nueva forma de expresar la idea de la Web semántica que, de forma entusiasta y vehemente, presentó **Tim Berners-Lee**, el inventor de la web, en los 16,2 intensos minutos que duró su charla del *TED*[7] en marzo de este año. La información que se encuentra a través de los documentos enlazados a través de las "viejas" tecnologías hipertextuales, esto es, la web indizable (o *crawlable web*) se ha interpretado en varias ocasiones como la punta del iceberg si se compara con la cantidad de datos que realmente se podrían encontrar en la Web.

"Primero se habló de orugas web *(webcrawlers)*, luego de arañas *(webspiders)*, ahora los motores de búsqueda son *dragones web*"

Las tecnologías semánticas, o mejor, el marcado semántico (xml, rdf, owl) ha progresado mucho en los últimos años, sobre todo a lo largo de 2009. Un conjunto de nuevas herramientas y técnicas, la participación de los motores de búsquedas más importantes, y el crecimiento de la nube de datos enlazados, han hecho que el mun-

do de la búsqueda en la Web (RII) haya cambiado, abriendo incluso un nuevo mercado.

Fue precisamente en el marco de la *Conferencia mundial de la Web* celebrada en abril de 2009 en Madrid cuando se afianzó la idea de los datos enlazados en la Web. A través de diversos trabajos científicos presentados en un workshop organizado, entre otros investigadores en la materia, por el propio **Berners-Lee**[1], se pudieron ver los servicios de información que conforman la nueva Web.

Ésta incluye *datasets* tan diversos y de dominios tan diferentes como: la *DBpedia* (**Bizer**, 2009), *Revyu*, *Geonames* para nombres geográficos, *BBC programmes*, bibliografía *Dblp*, datos de *Eurostat* en *Riese*, *Flickr*, *PubMed*, *UniProt*, así como múltiples datos de personas y sus relaciones definidos en vocabularios y servicios como *Foaf*, *Sioc*, *OpenCyc* o *Yago*, cuya disponibilidad permite crear aplicaciones para enlazar datos y reducir la dificultad a la hora de integrarlos.

La industria ha evolucionado rápidamente en esta simple idea desarrollando un conjunto interesante de aplicaciones, organizaciones y tecnologías. Los grandes servicios de búsqueda han entrado en una carrera de mejoras en sus sistemas para anotar, compartir, enlazar y recuperar datos. En 2008 fue *Yahoo!* el que lanzó *SearchMonkey*[2], una forma de personalizar (*customize*) la forma en que se visualizan los resultados mediante *plugins*. En mayo de 2009 ha sido *Google* el que ha dado el paso al anunciar sus recortes enriquecidos (*rich snippets*), que no es otra cosa que el reconocimiento y la presentación de metainformación embebida en el código fuente de las páginas.

"Necesitamos buscadores que permitan la búsqueda localizada (y también categorizada) a través de etiquetas en el ámbito de la web social"

La adopción de rdfa [el nuevo estándar del *W3C*[3] para utilizar los atributos de los elementos meta y link de xhtml] permite, como los microformatos, anotar semánticamente la información en un sitio/página web, en xhtml y, gracias a la evolución que ha experimentado rdfa en 2009, también en html.

"Anotar" es para los informáticos lo que para los documentalistas es asignar metadatos, o sea, enriquecer la información describiéndola a través de semántica. El 3 de junio de 2009 *Microsoft* lanzó *Bing*, denominado motor de decisión (*decision engine*), dando, a juicio de su director ejecutivo **Steve Ballmer**, un paso en la política de innovación en la búsqueda web, de tal forma que los usuarios puedan no sólo encontrar la información, sino también realizar tareas y tomar decisiones inteligentes.

Además de "los grandes" adoptando técnicas como rdfa u otras para el uso inteligente de la información, se unen este año nuevos actores en la escena de la búsqueda web: nuevos dragones web.

Éstos tratan de cubrir de misterio y glamour algo tan cotidiano ya para el usuario como la búsqueda. Los cuatro destacados durante 2009 son (**Sutter**, 2009): *Hakia* y *Twine*, que tratan de personalizar las búsquedas separando los resultados que pueden ser interesantes teniendo en cuenta preferencias o proyectos del usuario, al mejor estilo 2.0; *SearchMe*, que apuesta por una interfaz para que el usuario se mueva entre fotos e imágenes, en lugar de los tradicionales enlaces; y *Kosmix*, que agrupa la información por tipos para hacer más fácil su uso (por ejemplo, información de *Facebook*, de blogs, de la administración).

En mayo se lanzó además *WolframAlpha*, que tritura datos más que realizar búsquedas en el sentido tradicional de la RII. Sin embargo, si tengo que elegir los dos dragones web que han cautivado mi atención en este año, son: *Scoopler* (figura 1) y *Sindice* (figura 2).

Scoop quiere decir excavar con una cuchara, sacar provecho, e incluso también en el mundo periodístico, primicia. Esto es lo que hace *Scoopler*: excavar la información de la Web para sacar provecho de ella y encontrar primicias. Así, *Scoopler* no sólo busca en la Web, sino también en sitios dinámicos de *microblogging* como *Twitter*, o sitios para compartir enlaces como *Delicious*, lo que permite encontrar novedades y

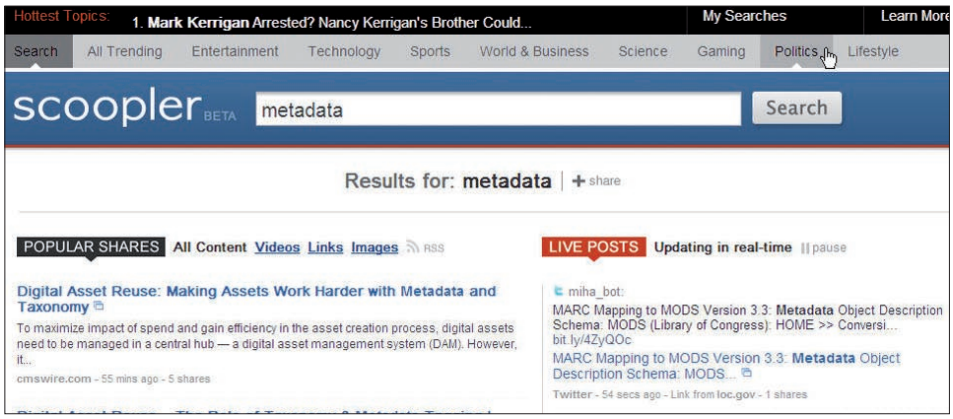


Figura 1. Búsqueda sobre "metadata" en Scoopler, http://www.scoopler.com

Figura 2. Búsqueda sobre information retrieval en Sindice, fuente: http://www.sindice.com

anuncios sobre un tema, agregando contenido a las búsquedas en tiempo real.

Sindice, de la mano de los grandes de la investigación de la Web semántica (**Stefan Decker**, del *Deri*, y **Nova Spivak**, creador de *Twine*), indiza rdf, rdfa y microformatos, permitiendo la búsqueda de datos específicos y también de propiedades de los recursos embebidas en ellos.

Scoopler y *Sindice* son los máximos exponentes de la búsqueda global en la Web en 2009. *Scoopler*, como dragón de la web 2.0, y *Sindice* como buscador por antonomasia de la web semántica. 2009 terminó con una gran variedad de productos, proyectos, conferencias y congresos, estándares, libros (**Croft**, 2009) y tesis (**Criado**, 2009) que, ratifican el espíritu –por fin– semántico de la recuperación de información: anotación, metadatos, vocabularios, inferencia, *tagging,* o simplemente, semántica.

3. Recuperación de información 2010: tendencias

Teniendo en cuenta todos los avances con los que hemos tipificado esa nueva generación de dragones web, tanto en la construcción de nuevos sistemas de recuperación de información como en el desarrollo de estándares para la información semiestructurada, podemos definir cuatro claves para 2010.

3.1. De web semántica a "datos enlazados"... incluso para el patrimonio digital

Desde un punto de vista heurístico, la búsqueda en la web semántica depende de las convenciones para representar cosas (por ejemplo, a través de ontologías, metadatos y otros vocabularios) y de la disponibilidad de los datos para utilizar esas convenciones. En general y salvo algunas excepciones como *Sindice* o *Hakia*, los buscadores semánticos no están dirigidos al usuario final. Sin embargo, la Web social ha cambiado, y seguirá haciéndolo, el estatus de disponibilidad de los datos, contagiando no sólo a los grandes sistemas de recuperación de información, sino también a los servicios de información digital (también conocidos como "bibliotecas digitales") que se generan en nuestro entorno más documental.

Los vocabularios son elementos constitutivos de la web semántica ya que proporcionan recursos terminológicos compartidos para la indización de la información web, el intercambio de datos y la integración de contenidos. Desde un punto de vista práctico, muchas de las aplicaciones de la web semántica se basan en ontologías ligeras y también últimamente en *Skos (Simple knowledge organization system)*, un modelo de datos que ha alcanzado el año pasado el nivel de recomendación del *W3C*, lo que implicará que a partir de ahora se desarrollarán servicios y aplicaciones que contemplen este estándar para la organización temática de la información, generando más datos codificados en rdf, susceptibles de ser recuperados, y enlazados.

Pero la Web de datos todavía adolece de datos. Para que la web semántica tenga el efecto que se le presume, tenemos que hacer lo que **Berners-Lee** (2009) instaba en su ponencia del *TED*, y lo que aprendemos de la web social: abrir los datos, abrir contenido cualificado al mundo de la web semántica, convirtiéndola en *LOD (linked open data).*

De la misma forma que los gobiernos están abriendo sus *datasets* (*data.gov*; *data.gov.uk*), otro ámbito natural para la apertura es el patrimonio cultural digital. Las bibliotecas y los servicios de información digital, que siempre han seguido con cierta suspicacia (por qué no decirlo) la web semántica, encontrarán en *LOD* un camino para la adopción definitiva de las tecnologías de ésta, haciendo por fin el patrimonio digital más accesible, usable y explotable. La misma *Europeana*, aún basada en su versión beta en algoritmos de búsqueda tradicionales, experimentará a lo largo de este año conexiones con la gran nube de *LOD*, enriqueciendo de esta forma la Web de datos.
http://www.europeana.eu

Estos son al menos los objetivos del ambicioso proyecto *Europeana versión 1*, de los cuales se contagiarán iniciativas como *Google books* que pronto considerarán la conexión de los *datasets* que extraigan en una Web más semántica.
http://version1.europeana.eu

Muchas otras bibliotecas, como la del *Cern*, han anunciado ya a principios de 2010 la migración de sus datos a datos enlazados.
http://www.cern.ch/bookdata

3.2. De metadatos a microformatos, rdfa, *Grddl* y microdatos

En la nueva definición de web semántica, web 3.0, o mejor, la denominación (otra vez extraña y afortunada de **Tim Berners-Lee**), *linked data*, necesitamos el desarrollo de nuevos estándares de codificación de metainformación, así como lenguajes de interrogación que permitan recuperar y extraer la información así estructurada. Los estándares clave en este sentido, serán rdfa, *Grddl* y los microdatos de html 5[5].

Rdfa es una recomendación del *W3C* para que los atributos de los contenidos web se expresen en datos estructurados rdf, en cualquier lenguaje de marcado, inicialmente en xhtml, pero también en html. Rdfa es la normativa alternativa del consorcio web, semejante a los microformatos. Los microformatos especifican tanto la sintaxis para embeber datos estructurados en documentos, como los vocabularios de cada microformato, rdfa especifica sólo la sintaxis, basándose en otras especificaciones de términos (vocabularios o taxonomías).

rdfa, que hizo en 2009 que *Google* creyera en esta web semántica "light", protagonizará a partir de 2010 distintos servicios de recuperación de información basados en los atributos de sus contenidos[4].

Grddl (*Gleaning resource descriptions from dialects of languages*, pronunciado algo así como "gridel"), es también una recomendación del *W3C*, de 2007, pero su utilización ha sido discreta hasta el momento. Sin embargo, se prevé que este estándar permitirá compatibilizar otros estándares con rdf, en el contexto de la búsqueda semántica para extraer distintos tipos de datos de documentos web.

Microdatos, la dimensión del nuevo y prometedor estándar del *W3C*, html 5[5], para anotar contenidos web con etiquetas legibles por máquina. Nuevamente la misma idea de pares atributo-valor, como en los metadatos "más tradicionales", para añadirlos a los documentos, en paralelo con los contenidos existentes. Html 5 es uno de los estándares más discutidos y proyectadamente más definitivos también, de la historia de la Web.

Microdatos es la forma de codificar metadatos en html 5, y que ha provocado que estándares semánticos como *Dcmi* (enero 2010) hayan abierto ya un debate sobre la forma de codificar su semántica descriptiva (datos *Dublin core*) en este nuevo estándar que, sin ninguna duda, dará que hablar en 2010, mezclando sus parámetros y asertos con rdfa para la nueva búsqueda web de datos enlazados.

3.3. De la *findability* a la búsqueda contextual y a la búsqueda de contenidos alternativos

Findability (buscabilidad o encontrabilidad) ha sido, por un tiempo muy razonable, el paradigma de **Peter Morville** (actualmente presidente de *Semantic Studios*) para definir la capacidad de un sitio, recurso u objeto de información digital de ser encontrado o recuperado tanto por los usuarios como por los sistemas de información de carácter global.

La búsqueda seguirá siendo clave en el futuro de servicios de información web, pero necesitamos migrar de búsqueda *booleana* a formas más sencillas (orientadas al usuario sin alfabetizar informacionalmente) para hacer búsquedas contextuales y/o alternativas.

Algunos ejemplos de estas nuevas formas de recuperar información incluyen: la utilización de metadatos para el cálculo de relevancia, la navegación facetada, las "*best bets*" (mejores apuestas), además de la búsqueda contextual. Un ejemplo de esta búsqueda contextual, es *HeadsUp*[3] (figura 3), un sistema de RII contextual y divertido, cuyo desarrollo se potenciará en el futuro próximo.

"La búsqueda migrará de la booleana a formas más sencillas orientadas al usuario sin alfabetizar informacionalmente"

Las bibliotecas y servicios de información digital necesitan mejorar las búsquedas federadas o capacidades de metabúsqueda, incorporando etiquetado social u otros tipos de descripción de información que fomenten y mejoren la búsqueda en el contexto de información, más allá de la *serendipity* que promueve, desde hace años, *Amazon*.

3.4. Del *page rank* a la búsqueda por etiquetas @,

Además de indizar la información automática o manualmente, de incluir metadatos formalizados y/o etiquetas libres, con microformatos u otras formas de codificación estructurada de las propiedades de la información, los datos web también se indizan de forma creciente con *hashtags*; esto es, acrónimos, palabras o simplemente

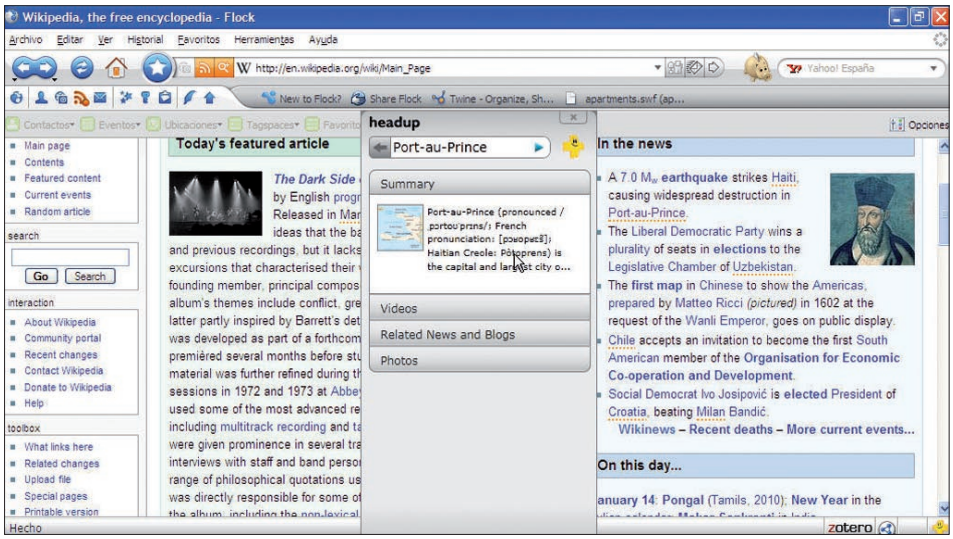


Figura 3. Heads-up (fuente: Wikipedia)

etiquetas temáticas que definen, una vez más, una convención dentro del ámbito de la Web, en este caso, de la Web social (ej. #okcon2010, o hashtag:#dcmi2010, etc.).

El *page rank* es el algoritmo que permitió que el cálculo de la relevancia en los motores de búsqueda como *Google* tuviese un cariz más cualificado. Sin embargo, éste no logra reflejar la naturaleza social del conocimiento en el entorno informativo dibujado ya para 2010. Por ello, necesitamos buscadores, sistemas RII que permitan la búsqueda localizada (y también categorizada) mediante etiquetas en el ámbito de la Web Social.

4. A modo de conclusión

Para mantener la Web de datos en el mismo estado saludable y creciente que hemos definido aquí, las herramientas de búsqueda del futuro deben basarse en lo que **Witten**, **Gori** y **Numerico** denominan una "diversidad descentralizada". Es necesario que la Web se interrogue de distintas maneras, permitiendo que coexistan diferentes indicadores de visibilidad, fortaleciendo una anarquía organizada de datos enlazados y potenciando la evolución que nos permite que incluso lo que pensamos (/creo que…) o lo que le decimos a otras personas (@evamen:…) sea una cantidad de conocimiento registrado y "buscable" sin precedentes. Sólo necesitamos darle paso al futuro, que está aquí.

5. Notas

1. *Linked data on the web: LODW Workshop*. Madrid, abril, 2009.
http://events.linkeddata.org/ldow2009/

2. *SearchMonkey*
http://developer.yahoo.com/searchmonkey/

3. Rdfa:
http://www.w3.org/standards/techs/rdfa#w3c_all

4. Por ejemplo, el *W3C* ha hecho público el 2 de febrero de 2010, un grupo de trabajo formal sobre rdfa, en el cual ponemos toda nuestra credibilidad de desarrollo.
http://www.w3.org/010/02/rdfa/

5. *Microdata* (discusión en html5).
http://www.whatwg.org/specs/web-apps/current-work/multipage/microdata.html#microdata

6. Un *plug-in* de *Firefox/Flock*, que permite ampliar la información contextual que aparece en recursos web, no sólo textual, sino también vídeos, fotos, etc. (en el ejemplo, Puerto Príncipe).

7. *TED (Technology, Entertainment, Design)*
http://www.ted.com/pages/view/id/5

6. Referencias

Berners-Lee, Tim. "On the next web". *Ted.com*, marzo 2009.
http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

Bizer, Christian et al. "DBpedia: A crystallization point for the web of data". *Journal of web semantics: science, services and agents on the world wide web*, 2009, n. 7, pp. 154–165.

Criado-Fernández, Luis. *Procedimiento semi-automático para transformar la Web en Web semántica*. Madrid: UNED, pp. 131 y ss. [tesis doctoral].
http://e-spacio.uned.es/fez/eserv.php?pid=tesisuned:IngInf-Lcriado&dsID=pdf

Croft, Bruce; **Metzler, Donald**; **Strohman, Trevor**. *Search engines: information retrieval in practice.* Addison-Wesley, 2009.

Manning, Christopher D.; **Raghavan, Prabhakar**; **Schütze, Hinrich**. *Introduction to information retrieval.* Cambridge: University Press, 2008.
http://informationretrieval.org/

Sutter, John D. "New search engines aspire to supplement Google". *CNN.com*, May 12, 2009.
http://www.cnn.com/2009/TECH/05/12/future.search.engine/index.html

Witten, Ian H.; **Gori, Marco**; **Numerico, Teresa**. *Web dragons: Inside the myths of search engine technology.* Elsevier, 2007.

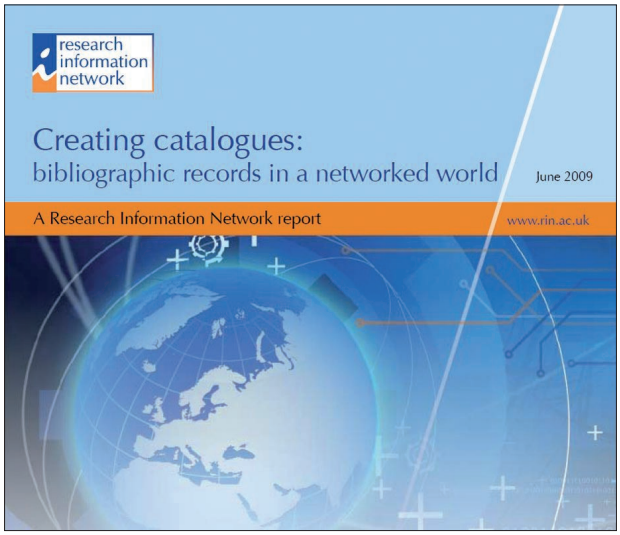
Informes anuales

Hacer catálogos en un mundo en red

Por **Lluís Anglada**

Creating catalogues: bibliographic records in a networked world. Londres: Research Information Network (RIN), June 2009, 48 pp.

Descargar el informe completo (2,3 MB), o resumen (*briefing*) (985 KB), o notas suplementarias (4,2 MB):
http://www.rin.ac.uk/creating-catalogues



Un colega reflexionaba sobre si los archiveros se habían equivocado dedicando tanto tiempo a debates normativos sobre cómo describir los documentos, y yo recordé algunas llamadas de atención en este sentido:

1. El informe *Rethinking how we provide bibliographic services for the University of California* (diciembre de 2005) de un grupo de trabajo sobre servicios bibliográficos de la *University of California*, que presenta varias recomendaciones sobre los cambios que estos servicios habrían de implementar para mejorar sus prestaciones:

Versión original inglesa (80 pp, 406 KB):
http://libraries.universityofcalifornia.edu/sopag/BSTF/Final.pdf

Traducción al catalán:
http://www.recercat.net/handle/2072/9103

2. El informe *On the record* de un grupo de trabajo de la *Library of Congress* sobre el futuro del control bibliográfico, enero de 2008 (49 pp., 448 KB)
http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf

3. El debate sobre la propiedad de los registros bibliográficos que se desencadenó cuando OCLC anunció una nueva política de re-uso de los registros de *WorldCat*:
http://bdig.blogspot.com/2009/07/sobre-oclc-proposito-del-ala-de-chicago.html

En junio de 2009 hemos tenido otra toque por parte de la *Research Information Network* (*RIN*), del Reino Unido, con su informe *Creating catalogues: bibliographic records in a networked world*.

Sitúa la catalogación en un nuevo contexto en el cual diferentes agentes proporcionan los mismos datos por lo que recomienda aprovecharlos más y hacerlos más visibles. Marca dos retos para los catálogos de hoy: ser eficientes (ahorro de recursos y más productividad) y aprovechar el nuevo entorno tecnológico de servicios basados en web.

Yo interpreto este informe a la manera de **Lorcan Dempsey**:

Catalogar cuesta demasiados recursos

No se aprovechan suficientemente los datos de otros agentes de la cadena. La catalogación ha funcionado a menudo bajo el principio de acercarse a las necesidades de los usuarios, pero esta plausible intención lo único que ha conseguido con seguridad es catalogar desde cero y con poca probabilidad de esta manera de servir mejor a nuestros clientes [estamos en 2010 y la *Biblioteca Nacional de España* no aprovecha la catalogación de, por ejemplo, la *Biblioteca de Catalunya*].

La catalogación está erróneamente enfocada a la diferencia y no a la similitud

Sus raíces están en la diferenciación de los ejemplares (estas diferencias son importantes para describir libro antiguo, por ejemplo) y no se han sabido agrupar las obras (ni con los títulos uniformes ni, por el momento, con los *Requisitos Funcionales de los Registros Bibliográficos*, *FRBR*). Los usuarios de los catálogos quieren encontrar juntos los diferentes ejemplares de la misma obra y esto nos lo ofrece *Library Thing* (p. ej.) y no en cambio nuestros catálogos.
http://www.ifla.org/files/cataloguing/frbr/frbr-es.pdf
http://www.librarything.com/

Los catálogos deben enseñar lo que tenemos

Y por esto tenemos que catalogar las colecciones especiales (tan a menudo escondidas) y procurar que nuestros registros formen parte de catálogos colectivos donde los usuarios los encuentren sin necesidad de que visiten el nuestro. A los usuarios no les interesa encontrar los documentos separados por tipos (libros aquí, artículos allá y documentos de archivo más allá).

Nuestra producción de metadatos debe poderse usar en nuestras herramientas, permitiendo que su re-utilización ahorre y enriquezca el trabajo de otros, y usando también nostros los de los demás. Quizá así nos llegue algún que otro usuario que por otros caminos no nos hubiera encontrado.

Esta nota se publicó el 4/10/2009 (en catalán) en:
http://bdig.blogspot.com/2009/10/catalogacio-en-xarxa-i-visible.html

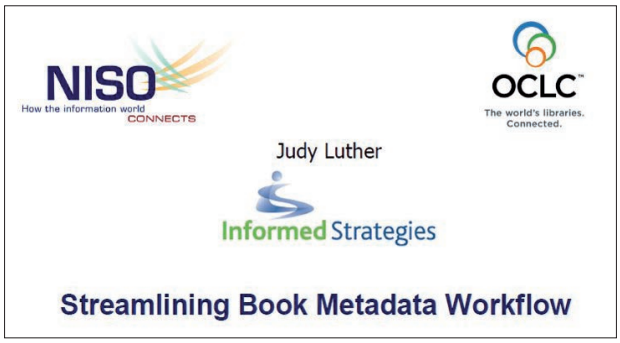
Libro blanco de *NISO-OCLC* sobre la creación y uso de metadatos en la cadena de suministro de libros

Luther, Judy. *Streamlining book metadata workflow*. A white paper prepared for the National Information Standards Organization (NISO) and OCLC Online Computer Library Center, Inc., June 30, 2009, 25 pp.

ISBN 978-1-880124-82-6

Published by:

NISO, One North Charles Street, Suite 1905, Baltimore, MD 21201 and OCLC, Inc., 6565 Kilgour Place, Dublin, Ohio 43017-3395



La *National Information Standards Organisation* (*NISO*) –Organización Nacional de Normas de Información de los Estados Unidos– y *OCLC* han anunciado la publicación de un libro blanco titulado Racionalización del flujo de trabajo de metadatos de libros.

Escrito por la consultora **Judy Luther** (de *Informed Strategies*), el documento analiza el estado actual de creación de metadatos, el intercambio y su uso en toda la cadena de publicación del libro. A través de entrevistas con más de 30 representantes de la industria, **Luther** ha hecho un mapa del intercambio de metadatos de los libros que ilustra el flujo de trabajo, y ha identificado oportunidades para la eliminación de redundancias y hacer todo el proceso más económico.

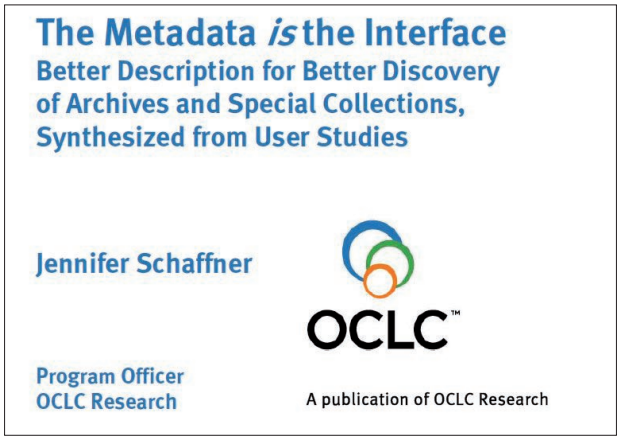
Nota de prensa original:
http://www.oclc.org/news/releases/200940.htm

Descargar el informe (1,44 MB):
http://www.niso.org/publications/white_papers/StreamlineBookMetadataWorkflowWhitePaper.pdf

Los metadatos son la interfaz: mejor descripción de archivos y de colecciones especiales

Schaffner, Jennifer. *The metadata is the interface: better description for better discovery of archives and special collections, synthesized from user studies*. Report produced by OCLC Research, 2009, 18 pp.

Descargar el informe (192 KB):
http://www.oclc.org/programs/publications/reports/2009-06.pdf
http://www.oclc.org/research/publications/library/2009/2009-06.pdf



Una treintena de años de estudios de usuarios enseñan que la precisión y la pertinencia son muy importantes para el descubrimiento en colecciones especiales, cosa que se evidencia sobre todo ahora que el descubrimiento ocurre en todas partes.

Desafortunadamente, existe una brecha histórica entre las expectativas de los usuarios y las prácticas descriptivas en los archivos y colecciones especiales. Hay que hacer cambios a la descripción porque los investigadores rara vez miran en los catálogos de bibliotecas o portales de archivos como primera opción. Garantizar que las "colecciones ocultas" puedan ser descubiertas requiere una descripción apropiada, no sólo el proceso que realizan los expertos, la catalogación y las búsquedas entre diferentes redes. Sería desolador si las colecciones especiales y los archivos permanecieran invisibles porque no tienen el tipo de metadatos adecuados para que puedan ser fácilmente descubiertos por los usuarios en la Web abierta.

En un artículo de 1986 sobre "El uso de los estudios de usuarios", **Bill Maher** describió a archiveros con instintos acerca de cómo se utilizaban sus colecciones (pero sin datos en los

que apoyar sus instintos) como "trabajando en la oscuridad". Desde entonces los estudios han demostrado de forma recurrente cuáles son las necesidades y preferencias de los usuarios cuando buscan en colecciones especiales y archivos. Con el tiempo los usuarios han adaptado sus tácticas de investigación: visitando los depósitos, mediante la consulta de catálogos impresos y guías, luego usando catálogos online y portales, y ahora a través de la Web.

Los estudios de usuarios han demostrado todo el tiempo que los metadatos descriptivos que informan de qué va un documento y su relevancia son de mucha importancia para el descubrimiento. Veinte años después ya no trabajamos en la oscuridad.

Catálogos en línea: lo que quieren los usuarios y los bibliotecarios

Calhoun, Karen; **Cantrell, Joanne**; **Gallagher, Peggy**; **Hawk, Janet**. *Online catalogs: what users and librarians want*. OCLC Report

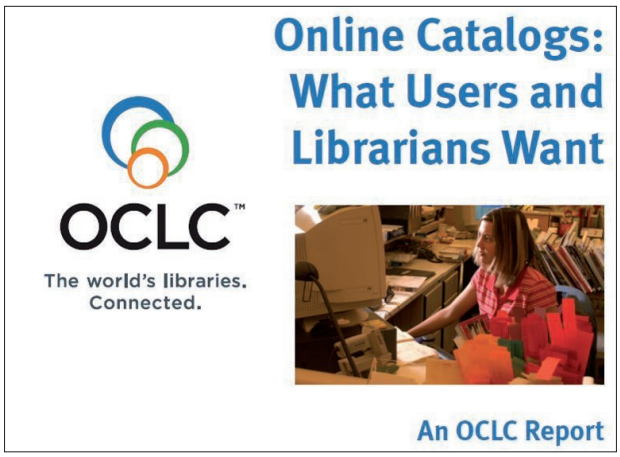
Dublin, Ohio. 3 March 2009, 68 pp.

OCLC Control Number: 311870930

ISBN: 1-55653-411-6

Descargar sinopsis en español, 24 pp., 202 KB:
http://www.oclc.org/reports/onlinecatalogs/213724lsb_Online_Catalogs_Synopsis.pdf

Descargar el informe (1,4 MB):
http://www.oclc.org/reports/onlinecatalogs/fullreport.pdf



El informe investiga sobre:

– metadatos más importantes para los usuarios para determinar si un material satisface o no sus necesidades;

– mejoras que los usuarios desearían ver en los catálogos online de las bibliotecas, y que les ayudarían a identificar el material adecuado;

– mejoras que los bibliotecarios recomendarían para los opacs que permitan facilitar su trabajo.

Los hallazgos indican, entre otras cosas, que si bien generalmente se piensa en los catálogos de bibliotecas como herramientas de identificación, para los usuarios la información sobre la disponibilidad de los documentos es de igual importancia.

Los resultados sugieren dos tradiciones de organización de la información en el trabajo: la de la biblioteconomía, y la de la Web. La concepción de los bibliotecarios acerca de la calidad de los datos sigue estando muy influenciada por los principios clásicos de organización de la información de su profesión, mientras que las expectativas de los usuarios finales de la calidad de los datos surgen en gran medida de sus experiencias de cómo la información se organiza en los sitios web populares. Lo que se necesita ahora es integrar lo mejor de ambos mundos en nuevas definiciones más amplias de lo que significa "calidad" en los opacs de las bibliotecas.

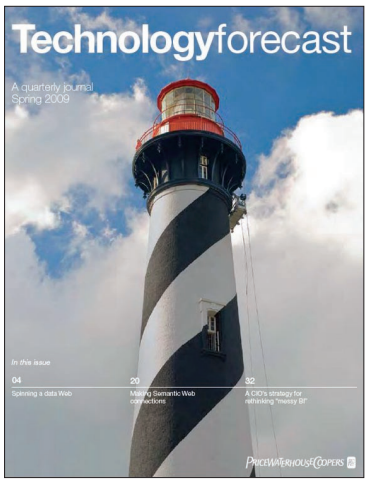
El informe concluye con recomendaciones para diseñar un programa de calidad de los datos equidistante entre lo que desean tanto los usuarios finales como los bibliotecarios, además de algunas sugerencias para futuras investigaciones.

Contextualización y web semántica

Technology forecast. Revista trimestral publicada por PriceWaterhouseCoopers. Spring 2009. Número dedicado a web semántica y ontologías.

La suscripción a *Technology forecast* es gratuita. Sólo hay que registrarse:
http://www.pwc.com/technologyforecast

Otros números recientes:



Winter 2010: *Unlocking hidden transformation value*.

Modelos de arquitectura de empresa, sistemas adaptativos, innovación.

Winter 2009: *Future of enterprise applications*

Summer 2009: *Technology trends in evergreen IT and cloud computing*. "Evergreen IT" define una futura situación ideal de flexibilidad, escalabilidad y drástica reducción de la complejidad de la tecnologías de la información.